Bridging logistic and OLS regression

Kapsalis, Constantine
Data Probe Economic Consulting

28. April 2010

DATA PROBE ECONOMIC CONSULTING INC.

9 Maki Place, Ottawa, Canada K2H 9R5
www.dataprobeinc.ca dataprobeinc@hotmail.com

BRIDGING LOGISTIC AND OLS REGRESSION

Constantine Kapsalis, Ph.D.

Working Papers Series: No. 2010-1

April 28, 2010

Abstract

There is broad consensus that logistic regression is superior to ordinary least squares (OLS) regression at predicting the probability of an event. However, OLS is still widely used in binary choice models, mainly because OLS coefficients are more intuitive than logistic coefficients. This paper shows a simple way of calculating linear probability coefficients (LPC), similar in nature to OLS coefficients, from logistic coefficients. It also shows that OLS coefficients tend to be very close to logistic LPC coefficients.

I. Introduction

There are several instances in economic studies where the dependent variable is not continuous but dichotomous (e.g. labour force participation, unemployment, poverty, reliance on social assistance). In these situations, the more familiar OLS regression has limitations and a logistic regression, or its very similar probit regression, is the appropriate choice. Specifically, the two main limitations of OLS are: (a) fitted values of *y* can fall outside the zero-one range; and (b) the error term *e* is necessarily heteroskedastic (Goldberger, 1964; Theil, 1981).

Unfortunately, logistic regression coefficients do not have the same intuitive interpretation as OLS coefficients do. In particular, in the case of OLS the dependent variable is the probability of the event itself (equation 1).

$$p = \beta_0 + \Sigma\, \beta_i X_i \tag{1}$$

In equation 1, p is the probability that the event will take place, and β_i is the partial derivative of p with respect to each X_i. For example, if the event is unemployment and X_i refers to the female gender, then the β_i coefficient shows how much more likely females are to experience unemployment than males, keeping all other attributes the same.

By contrast, in the case of logistic regression the dependent variable is not the probability of the event but its logistic transformation (equation 2).

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \Sigma\, \beta_i X_i. \tag{2}$$

Consequently, the β_i coefficients show the impact of each independent variable not on the probability of the event itself, but on its logistic transformation. The problem now is that, although the logistic model is more appropriate than OLS, we are left with regression coefficients that are difficult to interpret intuitively.

As a result, many practitioners recommend the OLS model as an approximation of the more correct logistic model or as a preliminary analysis tool (Moffit, 1999; Amemiya, 1981). This approach has been reinforced by the fact that the two models tend to lead to similar results, at least in terms of the partial derivatives of the dependent probability with respect to individual independent variables (Pohlmann and Leitner, 2003).

II. Logistic Linear Probability Coefficients

An alternative approach to relying on OLS is to derive linear probability coefficients (LPC) from the logistic coefficients. This way we can combine the superior statistical properties of logistic regression with the intuitive nature of OLS coefficients.

One approach that has been used to estimate LPCs is by comparing point estimates of the expected probability of various characteristics (Pohlmann and Leitner, 2003). For example, the LPC of the impact of female gender on the probability of unemployment can be derived from the results of a logistic regression by estimating the female and male probabilities, keeping the values of the rest of independent variables equal to their average value, and subtracting the two. Of course, since the relationship is non-linear, the

results will tend to differ depending on the choice of the point where the partial derivatives are estimated and the degree of non-linearity of the relationship.

The difficulty with the above approach is that it is computationally demanding. However, there is a simpler way of estimating LPCs from a logistic regression using the odds ratio. The odds ratio is a standard output of statistical packages, and it is simply the exponential value of the logistic coefficients. In logistic regression, odds are defined as the ratio $p/(1-p)$ and the odds ratio (Z) is defined as the ratio of two odds (equation 3).

$$Z = \frac{p_1 /(1 - p_1)}{p_0 /(1 - p_0)} \tag{3}$$

By solving the above equation for p_1 and assigning a specific value to p_0 we can easily estimate the corresponding LPC (equation 4).

$$\text{LPC} = p_1 - p_0 = (Zp_0 / (1 - p_0 + Zp_0)) - p_0 \tag{4}$$

In the case of dummy independent variables, p_0 will be the average probability of the omitted category. Using the previous example, in the case of gender the LPC will show the impact of being female on the probability of unemployment, keeping the rest of the rest of the female characteristics the same as those of males. In the case of a continuous independent variable (e.g. age) p_0 can be simply set equal to the overall average unemployment rate of the data sample.

III. An Example

We now present a simple example to illustrate the proposed methodology. The dependent variable is the probability of experiencing unemployment during the year among those who were in the labour force for at least part of the year. The independent variables include a continuous one (age) and several dummy variables (gender, education, province, area, and disability). The source of data is Statistics Canada's Survey of Labour and Income Dynamics (SLID), 2007. The sample includes 30,543 labour force participants, age 18-64.

Table 1 presents the standard SPSS regression results for OLS and logistic regression. The last column shows the LPCs of the logistic regression, based on equation 4 presented earlier. In addition to illustrating the method of estimating logistic LPCs, Table 1 reconfirms the finding in the literature that logistic and OLS regression results tend to be similar. In the case of the particular example, virtually all OLS coefficients were within one percentage point of the corresponding logistic LPCs.

IV. Conclusion

This paper has presented a simple way of estimating LPC from logistic regression results. It has also demonstrated with an example that OLS coefficients tend to be very close to logistic LPCs. Thus the paper provides analysts a simple way of combining the benefits of using logistic regression with the practical advantage of producing intuitive coefficients that are easier to communicate to a broader audience.

TABLE 1
OLS vs. logistic regression estimates of the rate of unemployment

	OLS		Logistic			
	b	t	b	t	Z	LPC
Constant	0.542	24.19	0.937	6.257	2.551	
Age (continuous)	-0.005	-31.02	-0.040	-29.63	0.960	-0.006
Sex						
- Male (omitted)						
- Female	-0.001	-0.204	-0.005	-0.172	0.995	-0.001
Education						
- Less than 9 years (omitted)						
- 9-10 years	0.020	1.355	0.114	1.132	1.121	0.019
- 11-13 years	0.039	2.582	0.097	0.955	1.101	0.016
- High school diploma	-0.054	-4.184	-0.394	-4.332	0.674	-0.056
- Some college	-0.031	-2.276	-0.264	-2.829	0.768	-0.039
- Some university	-0.021	-1.458	-0.210	-2.141	0.810	-0.031
- College diploma	-0.078	-6.404	-0.575	-6.669	0.563	-0.076
- University BA	-0.066	-3.737	-0.482	-3.662	0.618	-0.066
- University above BA	-0.108	-7.700	-0.909	-8.317	0.403	-0.108
Province						
- Newfoundland (omitted)						
- PEI	-0.014	-0.402	-0.091	-0.401	0.913	-0.018
- Nova Scotia	-0.055	-2.607	-0.318	-2.309	0.728	-0.059
- New Brunswick	-0.054	-2.450	-0.294	-2.032	0.745	-0.055
- Quebec	-0.088	-4.963	-0.515	-4.541	0.597	-0.091
- Ontario	-0.102	-5.793	-0.621	-5.507	0.537	-0.107
- Manitoba	-0.160	-7.806	-1.094	-7.605	0.335	-0.165
- Saskatchewan	-0.130	-6.197	-0.850	-5.878	0.428	-0.137
- Alberta	-0.151	-8.263	-1.013	-8.419	0.363	-0.156
- BC	-0.124	-6.862	-0.801	-6.771	0.449	-0.131
Area						
- Rural (omitted)						
- Urban: 0 to 29,999	-0.007	-0.783	-0.042	-0.645	0.959	-0.007
- Urban: 30,000 to 99,999	-0.017	-1.710	-0.112	-1.585	0.894	-0.017
- Urban: 100,000 to 499,999	-0.026	-3.016	-0.185	-2.944	0.831	-0.028
- Urban: 500,000 and higher	-0.010	-1.261	-0.066	-1.192	0.936	-0.010
Disability						
- No (omitted)						
- Yes	0.079	14.617	0.560	14.763	1.751	0.091

Note: The OLS R^2 was 6%; the logistic Nagelkerke R^2 was 9%.

References

Amemiya, T. (1981). "Qualitative response models: a survey", *Journal of Economic Literature* 19: 1483-1536.

Goldberger, A. (1964). *Econometric theory* (Wiley, New York).

Moffitt, Robert (1999). “New Developments in Econometric Methods for Labor Market Analysis”, in *Handbook of Labour Economics*, Volume 3, Chapter 24, Edited by O. Ashenfelter and D. Card.

Pohlmann, John T. and Dennis W. Leitner (2003). “A Comparison of Ordinary Least Squares and Logistic Regression”, *The Ohio Journal of Science*, 103 (5): 118-125.

Theil, H. (1981). *Principles of econometrics* (Wiley, New York).